BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 7 2005, pages 1138–1145
doi:10.1093/bioinformatics/bti133

Genetics and population analysis

Detection of low level genomic alterations by comparative genomic hybridization based on cDNA micro-arrays

Sven Bilke*,†, Qing-Rong Chen†, Craig C. Whiteford and Javed Khan

Oncogenomics Section, Pediatric Oncology Branch, Advanced Technology Center, National Cancer Institute, 8717 Grovemont Circle, Gaithersburg, MD 20877, USA

Received on September 3, 2004; revised on October 26, 2004; accepted on November 2, 2004
Advance Access publication November 11, 2004

ABSTRACT

Motivation: The accumulation of genomic alterations is an important process in tumor formation and progression. Comparative genomic hybridization performed on cDNA arrays (cDNA aCGH) is a common method to investigate the genomic alterations on a genome-wide scale. However, when detecting low-level DNA copy number changes this technology requires the use of noise reduction strategies due to a low signal to noise ratio.

Results: Currently a running average smoothing filter is the most frequently used noise reduction strategy. We analyzed this strategy theoretically and experimentally and found that it is not sensitive to very low level genomic alterations. The presence of systematic errors in the data is one of the main reasons for this failure. We developed a novel algorithm which efficiently reduces systematic noise and allows for the detection of low-level genomic alterations. The algorithm is based on comparison of the biological relevant data to data from so-called self–self hybridizations, additional experiments which contain no biological information but contain systematic errors. We find that with our algorithm the effective resolution for ± 1 DNA copy number changes is about 2 Mb. For copy number changes larger than three the effective resolution is on the level of single genes.

Contact: bilkes@mail.nih.gov

1 INTRODUCTION

Genomic alterations, such as gains or losses of specific DNA regions are frequently observed in tumors (Lengauer *et al*., 1998). The size of the affected regions can range between a few base-pairs (bp) to several mega-bp and may even cover whole chromosomes. Cancers of different diagnostic types have characteristic genomic alteration profiles (Lengauer *et al*., 1998), and some are predictive of aggressive behavior. Therefore, considerable efforts have been taken to map these genomic alterations for specific cancers in order to identify the genes responsible for the aggressive phenotype.

Metaphase comparative genomic hybridization (mCGH) was one of the first methods used to investigate DNA copy number changes which we refer to as 'genomic alteration' in this publication. Unfortunately, mCGH has several limitations including a relatively low spatial resolution (on the order of 10–20 Mb) and low sensitivity. The emerging methods of array-based comparative genomic hybridization (aCGH), which utilizes the BAC and cDNA microarray technologies, have overcome (Pinkel *et al*., 1998; Pollack *et al*., 1999) some of the problems associated with mCGH. When comparing sensitivities it has been demonstrated that BAC aCGH is more sensitive in detecting genomic alterations because of the larger length of BAC clones. However, the cDNA aCGH methodology has its own advantages. It has been successfully utilized to detect amplifications on the level of single genes. Because cDNA arrays can be used to measure DNA copy number as well as the transcript level, these arrays have also been successfully used to investigate the impact of gene dosage on the transcriptome (Hyman *et al*., 2002; Pollack *et al*., 2002). Many applications of cDNA based aCGH have focused on amplifications, where typically more than 10 extra copies of DNA are gained. However, low level DNA gains and losses are difficult to detect (Beheshti *et al*., 2003; Hyman *et al*., 2002) due to a lower signal to noise ratio. To increase the sensitivity for lower level DNA copy number changes, a 'running average' (RA) smoothing filter has been used as a noise reduction strategy (Hyman *et al*., 2002; Pollack *et al*., 2002). To our knowledge a systematic analysis of the sensitivity of cDNA-based array CGH has not been done. In our analysis we found that a significant part of the noise in cDNA aCGH data is of a systematic origin. However, the RA noise reduction strategy cannot effectively reduce this bias and therefore may lead to a low statistical significance. We have developed a novel algorithm which we named 'topological statistics'. This algorithm reduces systematic as well as statistical noise and allows the detection of low level DNA copy number changes with cDNA microarrays.

2 MATERIAL AND METHODS

2.1 Running average

The RA smoothing filter is a commonly used strategy for the reduction of stochastic noise. A basic assumption of this method is that the noise is independently and equally distributed over the individual probes of the cDNA array. In addition, it is assumed that the noise approximately follows a normal distribution. The measured relative intensity $D(x)$ for a probe at a genomic location x is therefore modeled as a linear combination of Gaussian noise $N(0,\sigma)$ with variance σ and the 'true' relative DNA copy number $S_j(x)$ multiplied by a response coefficient Γ:

$$D_j(x) = \Gamma S_j(x) + N(0,\sigma).$$

If the noise level σ is small compared to the signal ΓS it is sufficient to use a threshold Θ to define chromosomal gains by restricting

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

$\Gamma S > \Theta$.[1] The choice of the threshold Θ implicitly affects the false discovery rate α and false negative rate β. The complementary error function $\mathrm{erfc}(x)$ of the threshold, measured in units of variance, provides an estimate of these values,

$$\alpha = \mathrm{erfc}\left(\frac{\Theta}{\sqrt{2}\sigma}\right), \qquad \beta = \mathrm{erfc}\left(\frac{\Gamma S - \Theta}{\sqrt{2}\sigma}\right). \tag{1}$$

In cDNA aCGH data one often finds a signal to noise ratio $\Gamma/\sigma \approx 1$. For 'raw' data the cutoff-approach provides sufficient levels of significance only for relatively strong signals. For example, amplifications with typically $S \gtrsim 10$, can safely be detected. The detection of lower level gains and losses, however, requires the use of noise reduction strategies. The RA uses the fact that genomic alterations typically stretch over several neighboring sites. The true signal $S(x')$ is constant within some region $x' \in R$. Therefore averaging over a window of size W

$$D'_j(x) = \frac{1}{W}\sum_{i=-W/2}^{W/2} D_j(x+i)$$

reduces statistical noise if the whole window lies within R:

$$\sigma_r(W) = \frac{\sigma}{\sqrt{W}}. \tag{2}$$

The improvement of the signal to noise ratio is traded here for the loss of resolution: genomic alterations much smaller than the window-size W cannot be detected. Furthermore, the boundaries of altered regions are blurred as well. Therefore, one wants to choose W as small as possible but large enough to ensure the desired level of statistical significance parametrized by α and β. Combining equations (1) and (2), the smallest window size providing this significance level is given by

$$W = \left[\frac{(a+b)\sigma}{\Gamma S}\right]^2, \quad a = \sqrt{2}\mathrm{erfc}^{-1}(\alpha),$$
$$b = \sqrt{2}\mathrm{erfc}^{-1}(\beta). \tag{3}$$

If the size R of an altered region is smaller than the window size, the smoothened amplitude D' eventually falls below the detection threshold Θ. We define the size, where the expectation value of the smoothened amplitude equals the threshold Θ,

$$R_{\mathrm{struc}} = W\frac{a}{a+b}, \tag{4}$$

as the structural resolution. This is the minimal extension of a genomic alteration that can be detected at 50% of the chromosomal locations. Another resolution parameter of interest is the number of effectively independent measurements. It is obvious that the RA procedure strongly correlates the signal at adjacent sites. A measure of the minimal distance between effectively uncorrelated sites is given by the integrated autocorrelation time (Sokal, 1996),

$$\tau_{\mathrm{int}} = \frac{1}{2} + \sum_{t=1}^{\infty}\frac{\sum_x (D'(x+t)-\overline{D'})(D'(x)-\overline{D'})}{\sum_x (D'(x)-\overline{D'})^2},$$

where $\overline{D'}$ is the average of $D'(x)$ for all x. If the DNA copy number is constant in the range of summation one finds $\tau_{\mathrm{int}} = W/2$ for data smoothened by RA with window size W. The effective number $\tilde{N}$ of independent probes on the cDNA array which cover N genomic locations is therefore given by

$$N_{\mathrm{eff}} = \frac{2N}{W}. \tag{5}$$

2.2 Topological statistics

The application of the RA algorithm discussed in the previous section is based on several assumptions. Some of those assumptions are not met, more specifically that the noise is randomly distributed and that the variance is approximately constant on the whole genome. Our algorithm (Fig. 1) reduces this type of errors. Like the RA, it uses a sliding window to reduce statistical noise by combining multiple observations into one estimator. The difference to the RA is that this estimator is then compared to a null-distribution representing the unchanged DNA configuration. This distribution is obtained from a 'neutral' dataset, a measurement of a self–self hybridization experiment. In this type of experiments a single DNA sample is split into two groups, each of which is labeled with one of the two fluorescent markers respectively. Then the two groups are merged and hybridized onto one cDNA array. Naturally this experiment contains no biological data, but the systematic errors are similar to what is present in the measurements for biological samples. By comparing the two distributions a potential bias cancels out and ideally has no impact on the estimate. In order to increase the statistical significance several self–self hybridizations can be combined into one common distribution. Different from the RA, which makes a 'hard' assignment to either state 'normal' or 'altered', this algorithm provides a probabilistic estimate for the presence of genomic alterations and assigns a p-value to the center of the sliding window. In this paper we use mainly a t-test for the comparison of the two distributions, but the method is not limited to this statistics. For example we use also a Kolmogorov Smirnov test, which is independent of the normality assumption of the t-test. Most statistical tests do also take the noise level for the specific window location into account, thus reducing the impact of the inhomogeneous noise distribution.

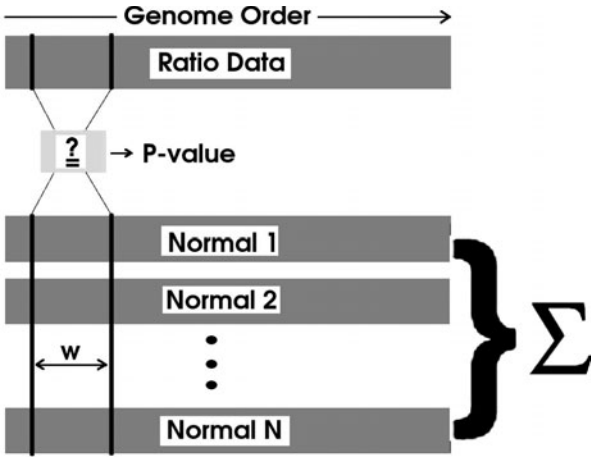

Fig. 1. Topological statistics: the distribution of observations in a sliding window of size W is compared to the distribution of normal data (copy number ratio 1) in a window at the same genomic location. Data from different reference datasets can be combined by considering the joint distribution. P-values for different distribution averages are assigned to the center of the sliding window.

2.3 Recurrent region

Recurrent genomic alterations, within the same disease, may play a role in tumorigenesis. It is hence possible that the genes encoded in these regions provide an advantage for the tumor cells when their copy number is changed. The identification of such regions is therefore highly relevant, and also the genes found in these regions may provide insights into the biology of the tumor and may constitute drugable genes.

Formally, a region is called 'recurrent' if the frequency,

$$\nu = \frac{\#\text{ samples with genomic changes}}{\#\text{ samples}} = \frac{C}{S}, \tag{6}$$

of its occurrence among tumor samples exceeds a certain threshold. When estimating this frequency, the noise present in the samples propagates to ν. In cases where the individual samples provide only weak significance for the

[1]The discussion for losses is straightforward and is left out for the sake of simplicity in what follows. For the same reason assume that the data was log-transformed, such that $<R> = 0$ for the regular DNA copy number.

presence of genomic alterations, a direct application of Equation (6) may lead to a systematic underestimation of ν due to an accumulation of type II errors. This effect can be reduced by considering the individual samples as 'repeated' measures of the same noisy variable 'presence of a *recurrent* genomic alteration'. Repeated observations, which individually may be only marginally significant can, as a whole, be highly significant. The average P-value

$$\bar{P} = \frac{1}{S}\sum_{i=1}^{S} P_i \propto 1-\nu, \tag{7}$$

combines S measurements into one estimator with a higher statistical power. If the null-hypothesis is true for *all* samples, the P_i are drawn from a flat random distribution and the expectation value is $<\bar{P}> = 0.5$. By virtue of the central limit theorem, in the absence of genomic alterations, the average of these P-values follows a normal distribution centered around 0.5 with variance $\sigma \propto 1/\sqrt{S}$. It follows that for a fixed significance level α the threshold γ for which the combined observation $\bar{P} < \gamma = 0.5 - \alpha\sigma$ can be called significant, increases with the number of samples.

When analyzing recurrent alterations, which are differentially affected in, for example, different stages of the disease, the frequency ν is of direct interest. Equation (7) states that $\bar{P}$ is an approximate estimator of ν. This can be seen as follows: The expectation value for a sample *without* a genomic alteration is $<P_\rho> = 0.5$, while for (marginally) significant samples it is $<P_\sigma> \approx 0$. Therefore the expectation value $<\bar{P}_{S,C}>$ for a set of S samples with C genomic changes is approximately

$$\begin{aligned} <\bar{P}_{S,C}> &= \frac{1}{S}\left(C<P_\sigma> + (S-C)<P_\rho>\right) \\ &\approx 0.5\frac{S-C}{S} = 0.5(1-\nu), \end{aligned} \tag{8}$$

proportional to the frequency ν.

2.4 Visualization

The visualization of the probabilities for gains and losses provided by topological statistics uses the frequently used color scheme. Red colors indicate a gain, while green indicates a loss at that location. The intensity

$$I \propto \begin{cases} 0, & \text{if } P > 0.05, \\ -\log(1/N) + \log(0.05), & \text{if } P \le 1/N, \\ -\log(P) + \log(0.05), & \text{otherwise}, \end{cases} \tag{9}$$

is proportional to the logarithm of the P-value. Points which are above a threshold $P > 0.05$ are shown in black. The P-values visualized in this way are *not* adjusted for multiple comparison, because we are asking for a genomic alteration at this specific site. A Bonferroni correction for multiple comparison is, however, built into the heat-map: the brightest intensity is clipped at $P = 1/N_{\text{eff}}$, where N_{eff} is the number of effectively independent clones [Equation (5)]. Consequently, if the numerically generated P-values perfectly followed the theoretical distribution expected for the null-hypothesis, one false positive finding with highest intensity is expected per array.

2.5 Data generation

2.5.1 Cell lines and genomic DNA We used four neuroblastoma cell lines in this study. The conditions for cell cultures were done as described previously (Khan *et al*., 2001). High molecular weight genomic DNA was extracted from interphase of a Trizol preparation for RNA extraction according to the manufacturer's instructions (Invitrogen, Gaithersburg, MD). Genomic DNA was treated with RNase A and protease (Qiagen, Valencia, CA), and purified by phenol/chloroform extraction followed by ethanol precipitation. We obtained normal genomic DNA samples (male, female or 1:1 mixture of male and female) from Promega, and genomic DNA samples containing the different numbers of X chromosomes (XXX, XXXX and XXXXX) from the NIGMS (http://www.locus.umdnj.edu/nigms/).

2.5.2 Microarray experiments Preparation of glass cDNA microarrays was performed according to a previously published protocol (Khan *et al*., 2002). Image analysis was performed using DeArray software (Chen *et al*., 1997). The cDNA library containing 42,000 clones was obtained from Research Genetics (Huntsville, AL) and clones were printed on two microscope glass slides as a set. Approximately 50% of the cDNAs on the microarrays were either known genes or similar to known genes in other organisms, whereas the remainder were anonymous ESTs. For aCGH experiments on cDNA microarrays, 20 μg of genomic DNA from neuroblastoma tumor or cell line samples were sonicated and purified with QIAquick PCR purification column (Qiagen, Valencia, CA). Three micrograms of sonicated DNA were labeled with aminoallyl-dUTP (Sigma) in a 25-μl reaction, including random hexamer (0.24 μg/μl, Roche), dATP, dCTP and dGTP (125 μM each), dTTP (25 μM), aminoallyl-dUTP (100 μM) and high concentration of Klenow fragment (2.5 U/μl, NEB). The labeling reaction was purified with QIAquick PCR purification column. Cy3 and Cy5 dyes were coupled to the reference DNA (1:1 mixture of normal male and female DNA) and sample DNA respectively. Cy3- and Cy5-labeled probes were then combined along with human Cot-1 DNA (50 μg, Invitrogen) and yeast tRNA (100 μg, Invitrogen). The mixture was concentrated and resuspended in 32 μl of hybridization buffer (50% formamide, 10% dextran sulfate, 4 × SSC and 2% SDS). The hybridization mix was first heated at 75°C for 10 min, then at 37°C for 1 h, and finally loaded to the pre-hybridized array. The hybridization was performed at 37°C overnight. The washing procedure was performed as described previously (Khan *et al*., 2001).

2.5.3 Data analysis Fluorescence ratios were normalized for each microarray by setting the average log ratio for each sub-array elements equal to zero (commonly referred to as 'pin-normalization'). The data was quality-filtered by removing those clones that had poor quality measurement (Chen *et al*., 1997) (quality <0.5) in more than 20% of all the samples. For the clones that passed this filter, the fluorescence ratio of low quality spot for the individual sample was replaced by the average ratio value of the remaining good measurements for that clone. The clones were then assigned to UniGene Cluster (April 2004). For the UniGene clusters represented by multiple clones, mean fluorescence ratios of those clones are used. After these processes we had 21 256 unique UniGene clusters remaining from the initial 42 591 clones. Map positions for the cluster were assigned by Blat searches against the 'Golden Path' genome assembly (http://www.genome.ucsc.edu/; July, 2003 Freeze). Throughout this publication, all genomic coordinates are given with respect to this assembly. Finally the clusters were sorted according to their starting position of sequence on each individual chromosome.

3 RESULTS

3.1 Running average

To test the sensitivity and resolution of cDNA aCGH to detect single copy number changes, we performed aCGH with DNA from cell lines containing different numbers of X chromosomes (1–5 copies) (Pollack *et al*., 1999) and compared them to a sample with two copies of X chromosomes. The autosomal chromosomes are normal for all these cell lines.

The observed mean fluorescence ratio of all clones across the X chromosome was calculated and is shown in Figure 2. The relatively large variance $\sigma \approx 0.12$ and the small response coefficient (regression slope) $\Gamma = 0.25$ make it difficult to detect the low level of DNA copy number changes. We used these numbers as estimates for the parameters in our theoretical analysis; with Equation (3) the *minimal* window size required to achieve reasonable statistical significance can be estimated (Table 1). We verified if the expected statistical significance is approximately observed when applying the RA with this window size W. The threshold was set to $\Theta = 1/2 * \Gamma S$, which together with W was chosen such that the *expected* false positive rate is $\alpha = 0.05$ and the false negative rate $\beta = 0.05$. Next we measured numerically the *empirical* false positive/negative rates.

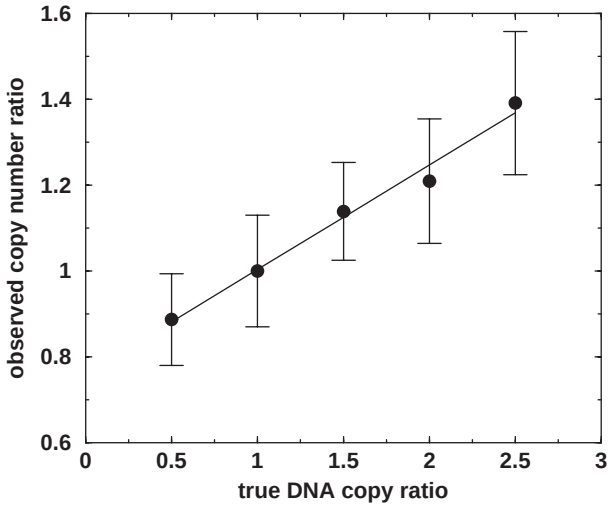

Fig. 2. The number of X-chromosomes divided by two is plotted against the average ratio observed in our CGH experiment. The error bars indicate the *standard deviation* in the data. The slope of the regression line is $\Gamma = 0.25$.

Table 1. Summary of the theoretical and numerical analysis of RA noise reduction

S	C	$\log_2 \Gamma S$	σ	W	α^*_{RA}	β^*_{RA}	α^*_{TS}	β^*_{TS}
0.5	1	−0.13	0.23	50	0.28	0.10	0.03	0.05
1.5	3	0.14	0.18	27	0.24	0.07	0.04	0.01
2.0	4	0.28	0.21	9	0.10	0.04	0.05	0.02
2.5	5	0.47	0.23	4	0.30	0.10	0.06	0.07
3.0	6	0.59^+	0.21^+	2	—	—	—	

Requiring, without adjustment for multiple comparison, significance levels $\alpha = 0.05$ and $\beta = 0.05$, the minimal window size W [Equation (3)] required to detect gain (losses) with copy number ratio S (copy number C for diploid cells) is estimated with the overfed standard deviation σ. The false positive/negative rates $\alpha^*_{\mathrm{RA}}, \beta^*_{\mathrm{RA}}$ observed with this window size and the threshold $\Theta = 0.5\Gamma S$ is listed. The values $\alpha^*_{\mathrm{TS}}, \beta^*_{\mathrm{TS}}$ were the corresponding rates obtained from TS with $p < 0.05$. Parameters marked with $^+$ were estimated by extrapolation of ΓS and using the average standard deviation of the observations.

The empirical false positive rate was estimated from the data of the autosome, where the copy number ratio is strictly equal to 1 in our benchmark. Conversely, the false negative rate was estimated from the X-chromosome data, where the true DNA copy number ratio is different from 1. Every observation a that satisfied $a > \Theta$ or $a < -\Theta$ was considered as a false positive in the autosome or on the X-chromosome a true positive gain or loss, respectively. We found that the observed false positive/negative rates were much higher than expected. While it was possible to increase the window size and change the cutoff Θ such that the 0.05 levels for the two rates was observed, the required window sizes were too large to get a reasonable resolution. For example we found empirically $W_{\min} = 200$ for the 1-copy loss. A window of that size covers almost $1/3$ of the entire X-chromosome.

3.2 Pathologies in cDNA array based CGH data

The RA strategy turns out to be much less efficient in detecting low-level gains and losses than anticipated by the theoretical analysis. The poor performance may be caused by the fact that some of the implicit assumptions of the RA strategy are not met to varying degrees. That is, the RA assumes that the noise in the data is identically and independently normally distributed. Non-normality is frequently observed in this type of data. Variance stabilizing transformations like the log-transform or generalized transformations (for two color microarrays, see for example, Rocke and Durbin, 2003) can improve this situation. In our log-transformed benchmark data most of the deviations from the normal distribution are in the low-ratio tail of the distribution (data not shown).The second assumption of an identical distribution is not met because the noise level strongly depends on the signal intensity. For example the noise level was found to be 50% higher in the region with a one-copy loss, the X-chromosome for male DNA, as compared to the autosome with two copies. But even in the autosome, with a constant DNA copy number, we observed a varying noise amplitude, which correlated with the GC content of the corresponding genomic region. The most basic assumption, namely that the noise follows a random distribution, is also not met. Systematic errors have been observed (Workman *et al*., 2002; Yang *et al*., 2002a) in microarray data and have also been linked to effects introduced in the commonly used DNA amplification protocol (Wilson *et al*., 2004). Furthermore cross-hybridization (Handley *et al*., 2004) is one source of systematic errors. For cDNA aCGH we found that the systematic bias is the biggest problem for the detection of low level genomic alterations. The scatter plot in Figure 3 demonstrates the non-randomness for two of our self–self hybridizations. This type of data does not by definition contain any biological information and should not contain any reproducible patterns. Therefore the observations should be uncorrelated and the points in the scatter plot should be distributed approximately spherically, given that the noise level in both measurements are similar. Instead we observed an ellipsoid with very different lengths of the two half-axes. The arrows in the diagram point to the direction of the eigenvectors (principal components) of the correlation matrix, and the length reflects the fraction of the variance explained by this component. The larger vector, which points in the direction of *reproducible* noise, accounts for 83% of the variability. In other words, the major fraction of the noise in the self–self hybridizations is of systematic origin. The reproducibility of the noise in the data alone does not fully explain why the noise scales so poorly in the RA smoothing filter. As long as the noise for the different measurements within a window is quasi-random, the scaling should be fine. However, we found that the low-frequency changes in the bias correlate with the low-frequency changes of the GC content on the genome. The correlation coefficient between the average GC content within a window including 200 cDNA spots and the average of the corresponding DNA copy number measurements in this window was $<\rho> = 0.6$. This indicates that the bias depends on the location of the genome, ultimately causing the poor scaling of the noise.

The third assumption, independence of the noise, is also violated: the systematic error at some chromosomal location should not depend on the signal at a different chromosomal location. In Figure 4 the distribution of correlation coefficients of the data with the true number Ξ of the X-chromosomes is plotted. Almost all the probes *on* the X-chromosome are strongly correlated, as expected. The average correlation coefficient is $<r_{\mathrm{X}}> = 0.77$. However, we also found a significant number of genes correlated with Ξ in the autosome. Due to the small number of degrees of freedom ($N = 5$ measurements per gene) a relatively broad distribution of correlation

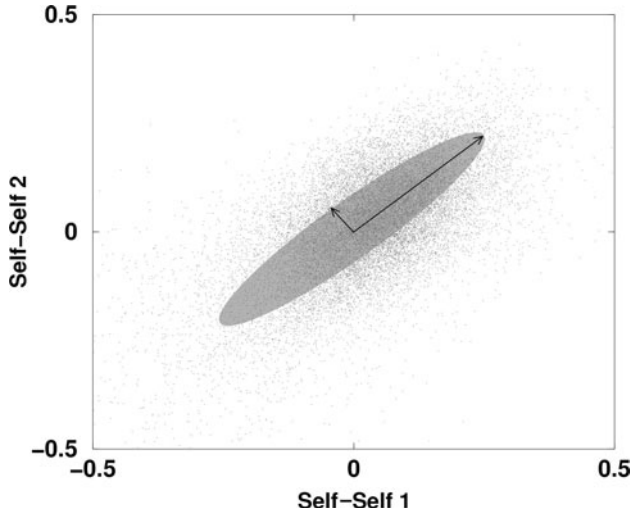

Fig. 3. Scatter plot comparing log-transformed expression ratios for two self–self hybridizations. The arrows point in the direction of the two principal components, and their lengths reflect the respective eigenvalues of the correlation matrix. The largest principal component points in the diagonal direction, indicative of strong systematic errors. The ellipsoid determined by the principal components covers 66% of the observations.

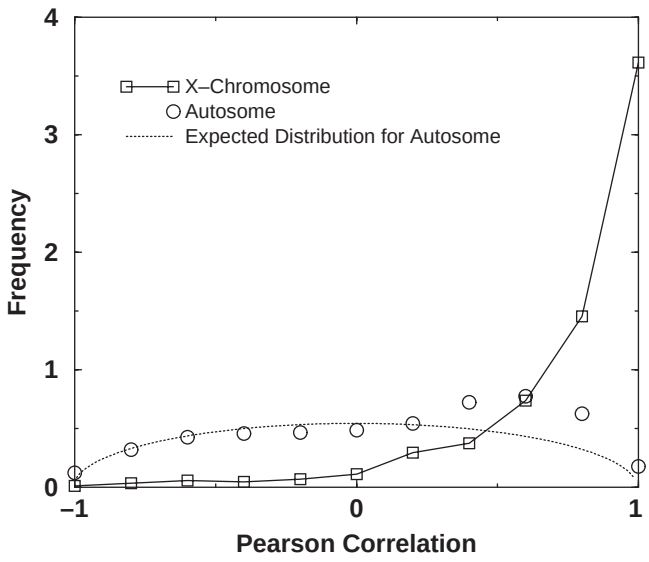

Fig. 4. The distribution of the correlation coefficients between the observed DNA copy number ratio and the true copy number of the X-chromosome. The solid line connects the frequency observations to guide the eye. The dashed line reflects the theoretical distribution (Press *et al*., 1992), Student's t-distribution with $N-2$ degrees of freedom and $t = r\sqrt{(N-2)/(1-r^2)}$. It represents the null-hypothesis that genes in the autosome do not respond to changes of the DNA copy number of the X-chromosome.

coefficients centered around $r = 0$ is expected. A few 'strongly' (anti)-correlated genes are expected by random chance. However we observed significantly more positively correlated genes than expected. This observation is reflected in the positive average correlation coefficient $<r_A> = 0.10$ and the 'bump' in the distribution around $r = +0.5$. The distribution expected theoretically is also shown in the same plot. The left part of the scaled distribution nicely fits the observation, while for positive coefficients we found considerably more correlated genes than expected. This result indicates that genes in the autosome respond to changes of the signal on the X-chromosome. A possible explanation for this behavior is cross-hybridization which was recently linked to unspecific binding of the poly(dT) in ESTs used for printing of microarrays (Handley *et al*., 2004). Alternatively our observation can be explained by errors in the clone annotation like the assignment to UniGene clusters.

3.3 Topological statistics

We developed a novel algorithm, topological statistics, to deal with the issues described above. The performance of this algorithm was tested on the same dataset as RA with equivalent parameters. A t-test was used for the comparison of the null-distribution $\mathcal{N}$ with the actual observations $\mathcal{D}$. This test calculates, like the RA, the average of the observations $\mathcal{N}, \mathcal{D}$ within the sliding window. Therefore the statistical noise scales similar to Equation (2). Consequently Equations (3)–(5), which estimate the required window size and the number of effectively uncorrelated clones, remain valid. We chose the same window size W used for RA benchmark for the numerical estimation of the false positive/negative rates α^*_{TS} and β^*_{TS}. A region was considered a 'genomic alteration' if the estimated P-value $p < \alpha = 0.05$, equivalent to the thresholds used in the RA benchmark. The observed false rates $\alpha^*_{\mathrm{TS}}, \beta^*_{\mathrm{TS}}$ are shown in Table 1. The results indicate that the noise is effectively reduced. Interestingly we found that the observed rates $\alpha^*_{\mathrm{TS}}, \beta^*_{\mathrm{TS}}$ are slightly better than expected for very low levels ± 1 copy number changes. We concluded that it is possible to identify low levels of genomic alterations when topological statistics is used to improve the signal to noise ratio. Topological statistics was designed to reduce the pathologies found in cDNA aCGH data, but it cannot reduce the noise introduced by cross-hybridization induced by genomic alterations elsewhere (here: in the X-chromosome) in the genome. This may explain why the observed rates $\alpha^*_{\mathrm{TS}}, \beta^*_{\mathrm{TS}}$ are larger than expected for the high level of DNA copy number changes.

As described earlier, a potentially present bias in the data is compensated in our algorithm by comparing data to a 'null-distribution' with the same bias, which therefore cancels out. To test the importance of this aspect of our algorithm we replaced the null-distribution $\mathcal{N}$ with random data drawn from a normal distribution with the same homogeneous variance. A major increase of the false discovery rate with this setup indicated that the reduction of systematic errors is a major aspect of our algorithm. For example, in the one-X-chromosome copy data the false positive rate increased to $\alpha^*_{\emptyset} = 0.17$. This is significantly larger than the $\alpha^*_{\mathrm{TS}} = 0.03$ when the self–self hybridization data was used for comparison. The observed $\alpha^*_{\emptyset}$ is still smaller than the $\alpha^*_{\mathrm{RA}} = 0.28$ found for the RA noise reduction. For all other copy number of the X-chromosome we found similar results. We attribute this to the fact that our algorithm still takes into account the non-homogeneous variance in the data, even when the background distribution $\mathcal{N}$ does not contain the information about systematic errors.

The other two artifacts, non-normality and the non-identical distribution, are also reduced, even though they turn out to be of less importance. The inhomogeneity variance is taken into account by almost all statistical tests one could use for the comparison of $\mathcal{N}$ and $\mathcal{D}$. However, the t-test we chose for this purpose is dependent on the

assumption of normality, which we found is also not met. We hence tried to replace the t-test by the Kolmogorov Smirnov test. It turned out that the lower statistical power of this test over-compensated the advantage of independence from the normality assumption. We observed increased α^* and β^* rates and therefore still used the t-test for all other experiments.

3.4 Neuroblastoma data

Neuroblastoma (NB) is one of the most common pediatric solid tumors, and accounts for 7–10% of all childhood cancers (Brodeur, 2003). The prognosis of patients with NB varies with stage and amplification status of the gene *MYCN*. Genomic alterations in NB have been investigated by cytogenetic and molecular methods including spectral karyotyping and metaphase comparative genomic hybridization. The most common genomic alterations observed in NB include loss of 1p36, gain of 17q and amplification in a neighborhood of *MYCN* on 2p25. Other recurrent changes including loss of 3p, 4p, 9p, 11q and 14q, and gains of chromosome 7 and 1q have been suggested to have relevance to the development and progression of NB.

As part of a larger study (Chen *et al*., 2004) we have performed cDNA aCGH analysis of four neuroblastoma cell lines for which genomic alterations have been characterized. For example the cell line SK-N-AS showing loss of DNA on 1p36 bounded by *CDC2L1* and *NPPA* has been reported (Cheng *et al*., 1995; White *et al*., 1995). In Figure 5 we show the output of our algorithm for the entire 1p36 cytoband for this cell line. Our data confirms the loss of DNA in a region inside the boundaries *CDC2L1* and *NPPA*. For comparison we also show the RA-smoothened data in the lower part of the figure. Additional to a much smaller region lost we find *gains* adjacent to the proximal boundary *NPPA* and close to the distal boundary *CDC21L1* when relying on RA. These gained regions are not reported in the literature and we find the same 'gains' in self–self hybridizations (data not shown). We therefore concluded that they are an artifact in our data, emphasizing that RA is less efficient in removing noise.

In Figure 6 we show a genome-wide analysis for four cell lines CHP134, IMR5, SMS-KCNR and SK-N-AS for which we found a characterization of the status for chromosome 17 in the literature. All cell lines were reported (Morowitz *et al*., 2003) to have a gain of 17q, which we confirmed with our analysis. We also confirmed the loss of DNA in 17p for SK-N-AS and the unchanged copy number for the other three cell lines. For comparison we also show heat-maps of the data before noise reduction and after application of RA. While RA increases the 'contrast' of the results, we find similar to the example above in the 1p36 region, several regions with 'false' gains or losses of genomic material. The heat-map for the TS filtered data is relatively clean; however, it contains several red and green lines spread over the genome. This is probably a reflection of the still relatively low signal to noise ratio even after removing systematic errors. The visualization used here does not fully correct for multiple comparison. For the brightest spots we expect one false positive line per array; for lower P-values represented by darker lines more false positives are expected. When considering multiple samples in order to detect recurrent regions, the increase of statistical power by the 'repeats' can compensate for this lack of statistical certainty. We demonstrated this in the rightmost panel, which shows the average P-value for gains and losses for all of the four cell-lines. In this panel one can see the recurrent loss of the distal arm of 1p and in 11q as well as the gain on 17q and 1q. The reported lower frequency of

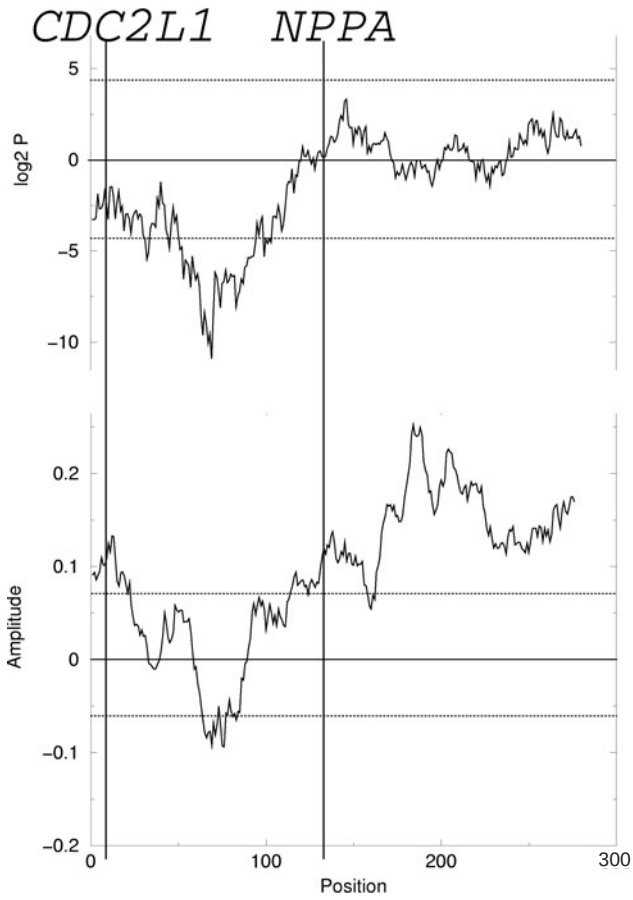

Fig. 5. $\log_2$ of the P-value for the presence of losses (negative values) or gains (positive values) estimated with TS for SKNAS (upper panel) in the 1p36 cytoband. The 5% siginficance levels are drawn with dotted lines. The two vertical lines indicate the location of the genes bounding the region described as lost (Cheng *et al*., 1995; White *et al*., 1995). The lower panel displays the same data smoothened with RA. The dotted lines indicate the threshold for 5% significance levels. The data was standardized such that the average log-ratio for the *whole* autosomal chromosomes is zero.

the recurrent gain of chromosome 7 (Stallings *et al*., 2003) is also visible. However, due to the small number of cell lines used in this study and the expected low frequency of this gain one finds that the observed signal is relatively weak.

4 DISCUSSION

Here we have shown that it is possible to detect the lowest ± 1 DNA copy number changes with cDNA aCGH when systematic errors in the data are sufficiently reduced. We analyzed both theoretically and experimentally the performance of the RA smoothing filter for cDNA array CGH data. We find that even under the assumption of 'well-behaved' noise, the minimal required window size needed to detect ± 1 DNA copy number changes is larger than the size 5 frequently used in the literature.

The sensitivity to detect low-level genomic alterations with cDNA aCGH was tested on a dataset with different numbers of copies of the X-chromosome and 22 pairs of autosomes. It turned out that the noise reduction performance of RA was much lower than expected

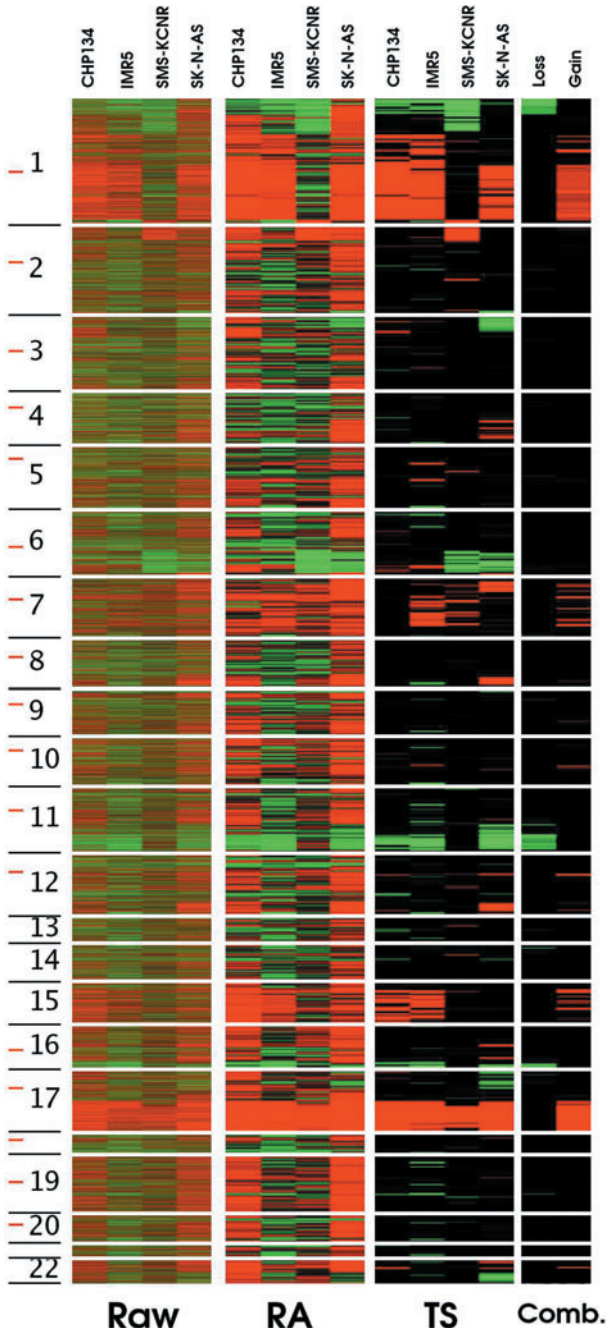

Fig. 6. Heat-map of the raw ratio data (RAW), the same data smoothened with RA and TS, both with window size 27 for four neuroblastoma cell lines (from left to right): CHP134, IMR5, SMS-KCNR and SK-N-AS. The right-most panel (Comb) visualizes the average P-value. Chromosomal location is displayed on the left with the location of the centromere marked in red.

due to systematic errors and other pathologies in the data. This result implies that it is impossible to detect a low level of copy number changes (± 1) with a reasonable resolution with this algorithm. While finalizing this manuscript, two algorithms for the analysis of CGH data were published, mainly focussing on an automatized detection of breakpoints. The authors (Haupe *et al*., 2004) report good performance of their algorithm for signal to noise ratio (SN) larger than $SN \approx 2.5$. This is considerably larger than $SN \approx 1$ present in our cDNA aCGH data for ± 1 DNA copy number changes. Unfortunately, for the second algorithm (Jong *et al*., 2004) no estimates for the effective resolution and sensitivity of their algorithm was discussed.

Our analysis of 'self–self' data demonstrated the importance of the systematic errors. The relatively strong correlation $<r> = 0.5$ between these supposedly uncorrelated datasets as well as the direction and amplitude of the principal components demonstrate that the major part of pathologies in the data is of systematic origin. In a related study (Yang *et al*., 2002b) self–self hybridization data was also used to assess the quality of array data. These authors focused on the level of noise, which they demonstrate is intensity-dependent. Consequently they suggested that intensity-dependent thresholds should be used for the detection of differentially expressed genes. In this study we have additionally demonstrated the presence of a systematic bias in the data and suggest an algorithm to reduce these bias' in CGH data.

In contrast to expression level measurements, the 'true' expected levels in DNA copy number experiments are often known and relatively easy to manipulate. This allowed us to estimate the sensitivity of the cDNA arrays and establish a benchmark for our algorithm. We could also directly observe the effect of cross-hybridization. Topological Statistics reduces this effect if the copy number of the cross-hybridizing DNA is unchanged because the same signal is present in the neutral dataset. However, if the DNA copy number is changed, the additional bias cannot be reduced by TS because this signal is not present in self–self hybridizations.

We demonstrated that TS reduces effectively both systematic and random noise. Like the RA, the reduction of statistical noise is performed by combining measurements within a sliding window into one estimator. The systematic errors are *canceled* out by comparing the window-content to observations of a neutral dataset, assuming that they approximately carry the same systematic errors. The effectiveness of TS strongly depends on this assumption; the best results were obtained when the self–self hybridizations used in the analysis were generated under similar conditions using the same print-batch.

The statistical significance obtained for the window sizes tested in this publication was slightly better than expected. For low-level changes, the statistical significance was sufficient to analyze small portions of the human genome. For a genome-wide analysis, the adjustment for multiple comparisons required stricter statistical thresholds. However, to compensate for these thresholds one may be forced to use very large window sizes reducing the resolution. Whole genome screening in high resolution for single copy gains and losses may be possible when focusing on the biologically relevant recurrent alteration regions. In this context, one naturally has to analyze multiple tumor samples which can be seen as independent samples for the presence (or absence) of genomic alterations at a specific location. The increase of the statistical power as a result of this 'repetition' eventually compensates for the adjustment for multiple comparisons. A possible problem arises from the fact that this procedure is very sensitive to the assumption that the samples are not biased. Even though we demonstrated that TS effectively reduces systematic errors, one cannot fully discount that an observation is caused by an undetected bias in the data. However, the prediction of a precise location allows one to use traditional high resolution methods like quantitative PCR to verify those regions. Without such

guidance it is, unfortunately, not practical to screen larger portions of the genome with these methods, due to their 'low throughput' nature.

Any sliding window smoothing method correlates measurements. It therefore reduces the spatial resolution of the microarray, which depends on the number of independent measurements. For a given window size, the number of effectively independent probes on the cDNA chip can be calculated by Equation (3) from the number $N \approx 21\,000$ of unique Unigene Cluster on our chips. With the simplifying assumption of an approximately homogeneous distribution of these cDNA clones on the human genome, consisting of approximately 3 Gb, the effective resolution of the array can be estimated. We demonstrated that one copy DNA changes ($c = +1$) can be detected with a window size of order $W \approx 27$. In regions encoding genes this implies roughly a 2 Mb resolution, which is significantly better than metaphase CGH and even BAC aCGH with a small coverage of BAC clones. We found that for the higher DNA copy number changes the required window size decrease quickly. For two copy changes ($\Delta S \geq 2$) the effective resolution is below 1 Mb. For $\Delta S > 3$ the required window size is 2. At this point the array reaches its maximum resolution; because of possible outliers it is common to consider a signal as trustworthy only if it is present in at least two consecutive probes. In other words, the cDNA aCGH platform can detect genomic changes at the *single* gene level (on average 160 kb) starting from four extra copies. To our knowledge this is among the highest resolutions of the currently available technologies.

REFERENCES

Beheshti,B., Braude,I., Marrano,P., Thorner,P., Zielenska,M. and Squire,J.A. (2003) Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization. *Neoplasia*, **5**, 53–62.

Brodeur,G.M. (2003) Neuroblastoma: biological insights into a clinical enigma. *Nat. Rev. Cancer*, **3**, 203–216.

Chen,Y., Doughherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomed. Optics*, **2**, 364–374.

Chen,Q.R., Bilke,S., Wei,J.S., Whiteford,C.C., Cenacchi,N., Krasnoselsky,A.L, Greer,B.T., Son,C.G., Westerman,F., Berthold,F. *et al*. (2004) cDNAs array-CGH profiling identifies genomic alterations specific to stage and MYCN-amplification in neuroblastoma. *BMC Genomics*, **5**, 70.

Cheng,N.C., Van Roy,N., Chan,A., Beitsma,M., Westerveld,A., Speleman,F. and Versteeg,R. (1995) Deletion mapping in neuroblastoma cell lines suggests two distinct tumor suppressor genes in the 1p35–36 region. *Oncogene*, **10**, 291–297.

Handley,D., Serban,N., Peters,D., O'Doherty,R., Field,M., Wasserman,L., Spirtes,P., Scheines,R. and Glymour,C. (2004) Evidence of systematic expressed sequence tag IMAGE clone cross-hybridization on cDNA microarrays. *Genomics*, **83**, 1169–1175.

Haupe,P., Stransky,N., Thiery,J., Radvanyi,F. and Barillot,E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Hyman,E., Kauraniemi,P., Hautaniemi,S., Wolf,M., Mousses,S., Rozenblum,E., Ringnér,M., Sauter,G., Monni,O., Elkahloun,A., Kallioniemi,O.P. and Kallioniemi,A. (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.*, **62**, 6240–6245.

Jong,K., Marchiori,E., Meijer,G., van der Vaart,A. and Ylstra,B. (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20** , 3636–3637.

Khan,J., Wei,J.S., Ringnèr,M., Saal,L.H., Ladanyi,M., Westerman,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 637–679.

Khan,J., Saal,L.H., Bittner,M.L., Jiang,Y., Gooden,G.C., Glatfelter,A.A. and Meltzer,P.S. (2002) Gene expression profiling in cancer. *Methods Mol. Med.*, **68**, 205–222.

Lengauer,C., Kinzler,K.W. and Vogelstein,B. (1998) Genetic instabilities in human cancers. *Nature*, **396**, 643–649.

Morowitz,M., Shusterman,S., Mosse,Y., Hii,G., Winter,C.L., Khazi,D., Wang,Q., King,R. and Maris,J.M. (2003) Detection of single-copy chromosome 17q gain in human neuroblastomas using real-time quantitative polymerase chain reaction. *Mod. Pathol.*, **16**, 1248–1256.

Pinkel,D., Segraves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C., Zhai,Y. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Pollack,J.R., Perou,C.M., Alizadeh,A.A., Eisen,M.B., Pergamenschikov,A., Williams,C.F., Jeffrey,S.S., Botstein,D. and Brown,P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.

Pollack,J.R, Sørlie,T., Perou,C.M., Rees,C.A., Jeffrey,S.S., Lonning,P.E., Tibshirani,R., Botstein,D., Borresen-Dale,A.L. and Brown,P.O (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci., USA*, **99**, 12963–12968.

Press,W.H., Teukolsky,S.A., Vetterling and W.T.m Flannery,B.P. (1992) *Numerical Recipes in C, 2nd edn*.

Rocke,D.M. and Durbin,P.M. (2003) Approximate variance stabilizing transformations for gene-expression microarray data. *Bioinformatics*, **19**, 966–972.

Stallings,R.L., Howard,J., Mullarkey,M., McDermott,M., Breatnach,F. and O'Mera,A. (2003) Are gains of chromosomal regions 7q and 11p important abnormalities in neuroblastoma? *Cancer Genet. Cytogenet.*, **140**, 133–137

Sokal,A. (1996) Monte Carlo methods in statistical mechanics: foundations and new algorithms. In DeWitt-Morette,C., and Cartier,P. and Folacci,A. (eds), *Proc. of ASI. Cargese, France*, p. 431.

Weaver,Z.A., McCormack,S.H., Liyanage,M., du Manoir,S., Coleman,A., Schrock,E., Dickson,R.B. and Ried,T. (1999) A recurring pattern of chromosomal aberrations in mammary gland tumors of MMTV-cmyc transgenic mice. *Genes Chromosomes Cancer*, **25**, 251–260.

White,P.S., Maris,J.M., Beltinger,C., Sulman,E., Marshall,H.N., Fujimori,M., Kaufman,B.A., Biegel,J.A., Allen,C., Hilliard,C. *et al*. (1995) A region of consistent deletion in neuroblastoma. *Proc. Natl Acad. Sci. USA*, **92**, 5520–5524.

Wilson,C.L., Pepper,S.D., Hey,Y. and Miller,C.J. (2004) Amplification protocols introduce systematic but reproducible errors into gene expression studies. *Bio. Techniq.*, **36**, 498–506.

Workman,C., Jensen,L.J., Jarmer,H., Berka,R., Gautier,L., Nielsen,H.B., Saxild,H.H., Nielsen,C., Brunak,S. and Knudsen,S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.*, **3**.

Yang,H., Dudioit,S., Luu,P., Lin,M.D., Peng,V., Ngai,J. and Speed,T.P. (2002a) Normalization for cDNA microarray data: a robust composite method for addressing single and multiple slide systematic variation. *Nucleic. Acids Res.*, **30**, e(15).

Yang,I.V., Chen,E., Hasseman,J.P., Liang,W., Frank,B.C., Wang,S., Sharov,V., Saeed,A.I., White,J., Li,J. *et al*. (2002b) Within the fold: assessing differential measures and reproducibility in microarray essays. *Genome Biol.*, **3**.